Korpus MetaFida

Tomaž Erjavec

Odsek za tehnologije znanja, Institut "Jožef Stefan"
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Mednarodni dan slovarjev 2022
ZRC SAZU
2021-10-12

Motivacija

- Korpusi na konkordančnikih CLARIN.SI so zelo raznovrstni, npr.
  - Referenčni: GigaFida
  - Znanstvena besedila: KAS
  - Starejša slovenščina: IMP
  - Uporabniško generirane vsebine: Janes
- Uporabniki bi radi iskali ali primerjali rezultate nekega iskanja po več korpusih, vendar:
  - je zamudno izvajati enaka poizvedbe po več korpusih
  - hitro lahko pride do napak
  - zaradi različnih oznak posameznih korpusov primerjave velikokrat sploh niso možne

MetaFida

- Izdelava združenega korpusa slovenskih korpusov na konkordančnikih (projekt RSDO)
- Izbira korpusov (=34 korpusov):
  - čim večje število čim bolj raznovrstnih korpusov
  - vključeni taki, ki niso podmnožica drugih korpusov
  - pojavnice označene vsaj z lemo in oblikoskladenjsko oznako
- Pretvorba korpusov:
  - identifikacija strukturnih in pozicijskih oznak
  - pretvorba v nabor oznak MetaFida
  - obdržimo samo tiste oznake, ki so skupne več korpusom
- Odstranjevanje podvojenih odstavkov:
  zbrisanih 12,5 % odstavkov in 6,5 % besedil
- Sortirano po letnici besedila (tisti brez letnice na koncu)
- = različica 0.1 korpusa MetaFida
- Različica 1.0 ob koncu projekta RSDO (nove različice korpusov, novi korpusi, druge spremembe)

Velikost in zgradba MetaFide

Pojavnic	4.463.244.126
Besed	3.646.106.520
Stavkov	228.797.305
Odstavkov	89.596.744
Besedil	15.338.751

- Strukturne oznake:
  text: corpus_id, corpus, info, id, year (-19.4 %), publisher (-3.8 %), title (-75 %), author (-99.4 %);
  p: id; **s**; **gap**; (**g**)
- Pozicijske oznake: **word**; **norm**; **lemma**; **tag_en**; **tag**
  + dinamični atributi lc; norm_lc; lemma_lc

Primer iz vertikalne datoteke

```
<text corpus_id="imp" corpus="IMP (starejša besedila)"
      info="https://www.clarin.si/noske/...corpname=imp"
      id="ZRC_00001-1584" title="Biblija (vzorec)"
      author="Dalmatin, Jurij" year="1584">
<gap/>
<p id="ZRC_00001-1584.p2">
<s>
INu         in          in          Cc        Vp
on          on          on          Pp3msn    Zotmei
je          je          biti        Va-r3s-n  Gp-ste-n
ſvoje       svoje       svoj        Px-nsa    Zp-set
dvanajſt    dvanajst    dvanajst    Mlc-pa    Kbg-mt
Iogre       jogre       joger       Ncmpa     Sommt
k'          k           k           Sd        Dd
ſebi        sebi        se          Px---d    Zp---d
poklizal    poklical    poklicati   Vmep-sm   Ggdd-em
<g/>
,           ,           ,           Z         U
```

Primer poizvedbe

NoSketch Engine | kostanj | metaFida v0.1 (združeni korpus)

Home
Išči
Seznam besed
O korpusu
My jobs
User guide

Shrani
Make subcorpus
Možnosti prikaza
  Usredinjeno
  Stavek
Razvrščanje
  po levi
  po desni
  iskani niz
  Podatki
  Premešaj
Vzorec
Filter
  Sub-hits
  1. zadetek v dokumentu
Frekvence
  Oznake niza
  Oblike niza
  Dokumenti

Iskalni niz kostanj 29,329 (6.57 na milijon)

Stran 1 od 978 Pojdi Naslednja | Zadnja

imp,WIKI00...	dobru sadene. O kreſu, kir je veliku	Kostajna	, je dobru stare Payni pomladiti,
imp,NUK_13...	, mandelne, leſhenke, orẹhe,	koſtan	, v'eno s'dobro poprej malu
imp,WIKI00...	. One jo nosijo is popovja divjih	kostanjov	, topolja ino drugega drevja. ɡ 72. K'
imp,WIKI00...	jablani, gruške, slive, tudi laški	kostanji	, ino proti konci totega meseca tudi
imp,WIKI00...	ſe. ɡ Sa lipo saſlushi predragi	koſtanj	pervo méſto, in she sato, ko to drevo
imp,WIKI00...	in gole proſtora s' divjim	koſtanjem	, kteri tudi na nar puſteji semlji
imp,NUKP14...	, snano je, de v kebrovim leti hraſt,	koſtanj	, oreh ino druge dreveſa bliso do
imp,NUKP14...	; per léſki, oréhu, hraſtu in	koſtanju	ſo tako imenovani brenklji ali
imp,NUKP14...	tovorizhi, kar je nam snano , ſo is	koſtanja	in nar vezhi is ſlibovne. ɡ
imp,NUKP14...	preš ali pa s posoljeno moko divjiga	kostanja	. ɡ Marnja. ɡ Na dveh sosednih njivah
imp,FPG_00...	kolerabami. ɡ Višnjev ohrovt s	kostanjem	. ɡ Peresa od štoržev odberi, čisto
imp,FPG_00...	, perdeni pečeniga in olušeniga	kostanja	, de se dobro skuha. ɡ Špargelnov
imp,FPG_00...	Tudi lahko popra in španske čebule,	kostanja	ali krompirja vanjo deneš in jo tako
imp,FPG_00...	drobnih, mandelnov in pečeniga	kostanja	vzame, tedaj se mora pa šest lotov na
imp,FPG_00...	na taljarčku in daj na mizo. ɡ Cukreni	kostanj	. ɡ Lepiga debeliga kostanja speci in
imp,FPG_00...	. ɡ Cukreni kostanj. ɡ Lepiga debeliga	kostanja	speci in olupi. Potem pa zavri cukra
imp,FPG_00...	. Vari, de cuker rujav ne postane.	Kostanj	pa na igle natakni, nekolikrat v
imp,FPG_00...	goveje mesa. ɡ Višnjev ohrovt s	kostanjem	in mesenimi klobasicami brez čev. ɡ
imp,NUKP14...	njih, bore, smreke, hraste, orehe,	kostanje	, jesene in breste izrediti, pri
imp,NUKP14...	driska napadla, stolci divjiga	kostanja	ali še bolje želoda v moko, in daj mu 2
imp,NUKP14...	to se po pravici mandeljni, orehi in	kostanj	prištevajo k redivni hrani. Mnoge

Primer sortiranja

NoSketch Engine

kostanj

metaFida v0.1 (združeni korpus)

Home
Išči
Seznam besed
O korpusu
My jobs
User guide

Shrani
Make subcorpus
Možnosti prikaza
  Usredinjeno
  Stavek
Razvrščanje
  po levi
  po desni
  iskani niz
  Podatki
  Premešaj
Vzorec
Filter
  Sub-hits
  1. zadetek v dokumentu
Frekvence
  Oznake niza
  Oblike niza

Iskalni niz kostanj 29,329 > Multilevel Sort 29,329 (6.57 na milijon)

Prva | Prejšnja Stran 362 od 978 Pojdi Naslednja | Zadnja Konkordance so razvrščene. Pojdi na: gos11

gfida20_de...	masla ali masti. Po nadevu potresi	kostanj	, ki mu primešaj za veliko prgišče
gfida20_de...	pošteno potruditi pri nabiranju	kostanja	, ki ga letos ni bilo ravno v izobilju
gfida20_de...	pomagali pri organizaciji in peki	kostanja	ter poskrbeli za prijetno druženje
gfida20_de...	v lepi septembrski soboti podal po	kostanj	na Blegoš, Milan Vošank pa po
gfida20_de...	Dejmo Stisnt Teater, pečenega	kostanja	, kuhanega vina in toplega čaja. Pri
gfida20_de...	pa s povabilom na prvi jesenski	kostanj	ter topel napitek. Programski del
gfida20_de...	gospoda Christiana Zaichena iz	Kostanj	na avstrijskem Koroškem. On vsako
gfida20_de...	Kostanje I Prijazna ta vasica je	Kostanje	, očaran vsak nad njeno je lepoto;
gfida20_de...	skupaj s koroškimi Slovenci iz	Kostanj	(Kőstenberg - Avstrija) na Triglav
gfida20_de...	, je letos obrodilo skupen pohod od	Kostanj	do Triglava. Prvo srečanje je bilo
gfida20_de...	strnil Christian Zeichen iz	Kostanj	na Avstrijskem Koroškem, ki je
gfida20_de...	borovnic, jeseni pa je precej tudi	kostanja	. Sredi gozdička nas je presenetil
gfida20_de...	gorami in petimi gozdovi je živel	kostanj	Kostanjček. Njegova starša sta
gfida20_de...	načrtujemo večdnevni pohod iz	Kostanj	(Avstrija) v smeri: Vrbsko jezero-
gfida20_de...	kaskada, ljubka jezerca, ogromni	kostanji	. Rozarij, velik vrt z
gfida20_de...	ponudili še krompirjeve svaljke s	kostanjem	v drobljencu vijoličastega
gos11,gos1...	greva kr na kostanjevo strjenko ne? ɠ	kostanji	ɠ ja evo ɠ upam da jih ne bo treba lupet ɠ ne
gos11,gos1...	kr pr pretresla ɠ če se ti razkuhajo	kostanji	je v bistvu dobr kr ti pol razpadejo
gos11,gos1...	sem bolj malo hodo rajš sem hodo po	kostanj	ker se je dobro prodajal ɠ a no vidiš
gos11,gos2...	pol vstanejo grejo malo nabirat	kostanje	je bla lih una štajon od kostanjev ne ɠ
gos11,gos2...	kostanje je bla lih una štajon od	kostanjev	ne ɠ ratajo lačni ɠ razumeš grejo malo

Zaključki

Zaključki

- Zelo na kratko predstavil korpus MetaFida v0.1
- Poleg osvežitve nabora korpusov so že predvidene izboljšave:
  - datum “notAfter” + sortiranje po njem
  - atribut lempos namesto lemma
- Če najdete napake ali imate predloge za nadaljnje izboljšave, prosim pišite na `tomaz.erjavecijs.si`!